A Hybrid Feature Selection Method for Effective Data Classification in Data Mining Applications

Ilangoan Sangaiya, K.L.N. College of Engineering, Madurai, India
A. Vincent Antony Kumar, PSNA College of Engineering and Technology, Dindigul, India

ABSTRACT

In data mining, people require feature selection to select relevant features and to remove unimportant irrelevant features from a original data set based on some evolution criteria. Filter and wrapper are the two methods used but here the authors have proposed a hybrid feature selection method to take advantage of both methods. The proposed method uses symmetrical uncertainty and genetic algorithms for selecting the optimal feature subset. This has been done so as to improve processing time by reducing the dimension of the data set without compromising the classification accuracy. This proposed hybrid algorithm is much faster and scales well to the data set in terms of selected features, classification accuracy and running time than most existing algorithms.

KEYWORDS
Classification, Data mining, feature selection, Genetic Algorithm, Symmetrical uncertainty

1. INTRODUCTION

Nowadays, across a wide variety of fields, a huge amount of data are being collected and stored in real-world databases at phenomenal rate. As the collected amount of stored information increases, the ability to understand and make use of it is not proportional. Also the users demand more sophisticated information. Therefore Feature selection in datamining helps to extract the relevant data from a huge data. Only a subset of relevant features out of all the available features is selected from the data being mined. While doing so the predictive accuracy of data mining algorithm improved by reducing dimensionality, removing irrelevant and redundant features. There are three general methods used in feature selection namely filter, wrapper and embedded. The filter approach preserves as much the relevant information as possible in the entire set of attributes without applying classification algorithm. Due to computational efficiency this method is quite popular even for large dataset with disadvantage being less computational effort and quality of selected features. Where as in wrapper method, attribute selection is done by taking classification algorithm and this is applied to selected attributes. This method selects attribute subset that is optimized for a given algorithm but it is too expensive for large dimensional data in terms of computational complexity and time. Finally, in the embedded approach the advantages of both approaches are used by implementing the diverse evaluation criteria in different search phases. The embedded approach is capable to achieve accuracy of a wrapper method at the speed of filter method.

DOI: 10.4018/IJGHP.2019010101

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
2. RELATED WORK

A lot of recent research works related to hybrid feature selection are being proposed some of the works only referred in this section. Ron Kohavi (1997) proposed a wrapper technique for feature subset selection by explored the relation between optimal feature subset selection and relevance. They also had an exploration of relation between optimal feature subset selection and its relevance using the wrapper technique for feature subset selection. In 1999, Mark A. Hall has proposed Correlation based Feature Selection (CFS) (Hall, 2000). In his work he approaches feature selection for machine learning through a correlation based approach. Ron Kohavi along with Dan Sommerfield (1995) used wrapper selection method. This method includes induction algorithms such as holdout, bootstrap and cross-validation for estimating the prediction accuracy. Using genetic algorithm a multi-criteria optimization problem of feature subset selection was done by Jihoon Yang and Vasant Honavar (1998). Pengyi Yang and Zili Zhang (2009) proposed a hybrid algorithm named GAEF. In this approach, there are two stages under feature selection process. In the first stage, GA is employed for preselecting the features and in next stage a small feature subset for accurate sample classification is identified using filter selector. Besides this work they (2010) also Proposed Multi-Filter enhanced Genetic Ensemble (MF-GE) system. Hybridizing the ReliefF, mRMR filter and GA a three-stage selection algorithm was proposed for gene selection problem by Shreem Salam salameh and Abdullah Salwani (2013). In Minimum Redundancy Maximum Relevance (mRMR) and GA was combined and proposed as a two stage feature selection algorithm by El Akadi, A. Amine, A. El Ouardighi, and D. Aboutajdine (2009). Lei Yu and Huan Liu (2003) have proposed a new algorithm which meets the demand for feature selection for high dimensional data. Payam Refaelzadeh, Lei Tang and Huan Liu (2007) have evaluated three Feature Selection (FS) algorithms: ReliefF, FCBF and IG and also compared the performance of decision tree (C4.5), Naive Bayes (NBC), Nearest Neighbor (NN), Support Vector Machines (SVM). A lot of efforts have also been made in the area of GA based feature selection. Feng Tan, Xuezheng Fu, Yanqing Zhang, and Anu G. Bourgeois (2007) proposed a two stage method. In the first stage several feature selection methods are applied on a original feature set and then GA is used to find the optimal feature subset. The GA based SVM wrapper feature selection method through properly designed chromosome and fitness function for hyper spectral data is proposed by Li Zhuo (2008). An attribute selection using Genetic Algorithm is briefed by M. A. Jayaram (2010), along with four different classifiers GA is used as random search method and Radial basis function is used as induction method wrapped with Genetic Algorithm. J. Zhou et al. (2008) proposed an Ant Colony Optimization and Mutual Information based feature selection for equipment fault diagnosis. In which Regression estimation model and mean squared error are used for the feature subset evaluation. Chun-Kai Zhang et al. (2005) proposed Ant Colony Optimization and Mutual Information based hybrid feature subset selection algorithm for weather forecasting. A fuzzy-rough data reduction using any colony optimization and C4.5 was proposed by R. Jenses et al. (2005). Xiangyang Wang et al. (2007) proposed a rough set based feature selection using PSO. H. Liu et al. (1996) proposed a consistency based feature selection mechanism it finds the value of the attributes using the level of consistency in the class values when the training instances are projected onto the subset of attributes.

3. PROPOSED METHOD

3.1. Symmetrical Uncertainty

The symmetrical uncertainty between features and the target concept can be used to evaluate the goodness of features for classification. The feature with a larger SU value gets higher weight. SU measures the correlation between two variables X and Y based on the information theory. It can be calculated as follows.
Modeling and Analyzing of Research Topic Evolution Associated with Social Networks of Researchers

The Financial Clouds Review
[www.igi-global.com/chapter/financial-clouds-review/64529?camid=4v1a](www.igi-global.com/chapter/financial-clouds-review/64529?camid=4v1a)